图书情报工作
第 66 卷 第 6 期 2022 年 3 月

基于决策树的多源文献元数据融合研究*

■ 李静 胡潜 李想 肖兵
华中师范大学信息管理学院 武汉 430079

摘 要：[目的/意义]构建多源文献元数据融合模型，有助于提升文献元数据整体质量，促进资源发现系统中的元数据管理与利用，优化用户资源发现服务体验。针对笔者此前提出的文献元数据判重策略进行优化，从经验为主向自动化转变，在保障判重和融合效果的前提下，提升整个过程的自动化水平。[方法/过程]针对不同类型文献的元数据项不一样、同一文献不同来源的元数据项不一样均会使得判重方法有所区别的情况，提出一种自动化的基于决策树的多源文献元数据融合模型，将判重问题转化为分类问题，根据特征相似度选择特征并构造决策树，在此基础上实施元数据判重及融合，并以不同类型的文献资源元数据为例进行实验，对策略进行效果验证。[结果/结论]结果显示，对于 5 种文献类型元数据，判重策略的准确率均达到 99% 以上，召回率均达到 98% 以上，总体效果较好。对于融合策略的效果判断，专利、学位论文、期刊论文、会议论文、图书的元数据项质量提升比例分别为 15.15%、36.80%、15.29%、52.63%、15.38%，均有明显幅度的提升。

关键词：多源元数据 决策树 元数据判重 元数据融合
分类号：G254
DOI：10.13266/j.issn.0252-3116.2022.06.013

1 引言

大数据环境下，各类型文献资源爆炸式增长，传统图书馆难以应对复杂多变的文献资源发展需要，而资源发现系统的日益成熟促进了各类型文献资源的开放融合与协作共享。在资源发现系统中，元数据的重要性不言而喻，一方面能够帮助用户快速获取所需资源，另一方面对海量元数据进行抽取、映射及导入等处理，能够有效避免不同来源、资源及数据结构带来的组织障碍，辅助服务商提升系统管理服务水平。但是各资源发现系统中也存在诸如元数据加工成本及系统适应性导致服务商提供的元数据厚薄不一，质量参差不齐，标注过程出现错误，系统间互操作性较差等问题。单一来源的元数据无法解决上述问题，因此，需要对多源文献元数据进行融合重组，从元数据准确性、完整性等方面全面提高资源发现系统中的文献元数据质量，提高元数据管理和利用水平，充分发挥资源发现系统的发现服务组织价值，从而提升用户体验。

为解决上述元数据质量问题，林鑫等[1]研究借鉴判重和融合思想，提出了基于多源数据融合的文献元数据质量提升模型。实验表明，对于期刊论文元数据，该策略的准确率为 99.9%，召回率为 99.2%[1]，策略的整体效果较优。然而，不同类型文献的元数据项不一样，其可用的判重方法会不同；同一文献不同来源的元数据项不一样，其实际采用的判重方法可能也会有所区别。而上文中期刊元数据项判重策略主要以经验选择为主，这就导致该方法的普适性不足。针对这些问题，本文构造了自动化的基于决策树的多源文献元数据融合模型，将判重问题转化为分类问题，采集多来源的文献元数据并进行预处理，根据元数据项构造特征，通过计算特征相似度进行特征选择并构造决策树，在此基础上实施判重，形成待融合元数据，最后通过对待融合元数据的融合处理，最终生成对文献资源准确、一致、完整的描述。

2 相关研究

元数据作为图情领域研究与实践的重要主题，近年来受到了国内外相关领域学者的广泛关注，与本文相关程度较高的研究包括元数据质量评价指标、元数据质量控制和元数据融合 3 个方面。

* 本文系国家社会科学基金项目"'互联网+'背景下面向产业链的行业信息服务融合研究"（项目编号：16BTQ063）研究成果之一。
作者简介：李静，博士研究生，E-mail：lj2016122579@mails.ccnu.edu.cn；胡潜，教授，博士生导师；李想，硕士研究生；肖兵，博士研究生。
收稿日期：2021-08-01 修回日期：2021-10-15 本文起止页码：118-125 本文责任编辑：杜杏叶

对元数据质量进行评价是开展质量控制和融合的前提。国外研究中,J. R. Park 等对全美元数据相关从业者进行随机问卷调查,结果显示从业者普遍认为完整性、准确性和一致性是影响元数据质量最基本且最深远的 3 个指标[2];B. Stivilia 等在对信息质量变化原因探索的基础上,提出完整性、准确性、一致性和复杂性等评价指标[3];T. R. Bruce 等在对 B. Stivilia 等提出的指标因素框架进行凝练后,将元数据质量评价指标扩充为完整性、准确性、期望满足程度、一致性、可用性、时效性和来源 7 个方面[4]。当前国内研究主要从通用系统和特定系统两个角度对元数据评价指标展开工作。通用系统评价指标研究中,相关学者将其归纳为完整性、准确性、可获得性、可理解性、易用性、合规性、及时性、一致性、开放性和客观性等[5-7]。特定系统评价指标研究中,张晓娟等提出政府数据开放平台元数据质量评价指标包括存在性、一致性和开放性[8];董微等结合当前图书馆发展现状,将开放期刊元数据质量指标划分为准确性、完整性、及时性、唯一性、一致性、有效性和关联性 7 个方面[9];刘家真等认为电子文件管理元数据评价指标包括描述程度、描述精度、数据现时性 3 个方面[10]。

通过对元数据质量控制研究成果的调研发现,相关学者认为元数据质量问题主要包括著录规范性不足、准确性不足、完整性不足、厚度不足、元数据重复、元数据判重错误等方面[1,9,11]。针对上述元数据质量问题,学者们也开展了相应的元数据质量控制研究,包括元数据清洗、映射、判重及分阶段过程控制几个核心环节。G. L. Li 等指出为了保障综合效果,清洗环节应重点关注数据质量、时间效率和成本控制[12];李慧佳等研究了 WoS、EI、CNKI 及 CSCD 4 个来源的机构名称元数据的语义化映射策略[13];在元数据判重环节,相关学者主要是围绕元数据取值相等的思路开展研究[14-15];在分阶段过程控制中,H. Manguinhas 等介绍了 UNIMARC 书目元数据模式,并以 XML 格式对元数据进行质量记录,从而实现质量控制过程的自动化[16];曹月珍等指出制定元数据质量控制标准应统揽全局,从制定元数据标准、元数据加工、系统录入、更新、系统间互操作、元数据评估等各阶段入手,进而实现全过程控制[17]。

资源发现系统中元数据结构和内容等存在的质量问题会导致系统间互操作性较差,因此需要通过多源元数据融合解决上述问题。王利亚等针对健康医疗数据中的元数据多源、多样、分散及非结构化等问题,利用 NLP 算法对数据进行去重、归一和消歧等,并构建了健康大数据平台[18]。严承希等认为元数据是非规范性的输入单元,需要对其进行转换与复合,提出的冲突冗余检测思想及知识合并方法为本文的元数据判重及融合部分提供了参考[19]。

综上,目前关于元数据质量评价指标、质量控制和元数据融合的相关研究业已成熟,学者们提出的质量评价指标相似度较高,主要为准确性、一致性和完整性 3 点,在具体的应用实践中会有些许差异。元数据质量控制主要是从元数据规范化、映射、判重及过程控制管理等方面展开。根据元数据融合研究可知元数据存在的问题包含不全面、准确性不足、非结构化、非规范化等。因此,针对元数据存在的问题,本文借鉴元数据通用评价指标和质量控制思想及元数据融合方法,以获取准确、一致和完整的元数据为目的,对已有元数据判重策略进行改进,允许元数据取值部分相等,在保证准确率的同时提高召回率,扩大多源元数据融合比例,尽可能保障文献资源的准确性、一致性和完整性。

3 基于决策树的多源文献元数据融合模型

为了得到准确、一致、完整的文献资源,需要获取尽可能多的优质元数据对其进行描述,本文根据林鑫等[1]研究中元数据融合障碍分析,设计了基于决策树的多源文献元数据融合模型。该模型主要包括 4 个模块:多源文献元数据采集、元数据预处理、特征选择及元数据判重、元数据融合,模型如图 1 所示。首先,多源文献元数据采集模块的采集对象是多来源文献元数据,其是元数据判重及融合的来源;其次,对采集到的多来源元数据进行规范化处理,为后面的判重与融合提供便利;再次,通过元数据项构造特征及计算特征相似度,然后进行特征选择并构造决策树,进而实施判重处理;最后利用元数据融合模块对待融合元数据进行融合,最终生成准确、一致、完整的元数据。

3.1 多源文献元数据采集

为了获取准确、一致、完整的文献资源,实现资源发现系统跨平台的一站式检索,需要对资源发现系统中的元数据进行判重与融合。而资源发现系统中的文献元数据来源于多个文献资源数据库,如中国知网、万方数据库、百度文库、微软学术搜索、Web of Science 等,因此,本文选取不同的知识服务平台作为文献资源元数据来源库, 采集不同数据库中多种类型的文献元

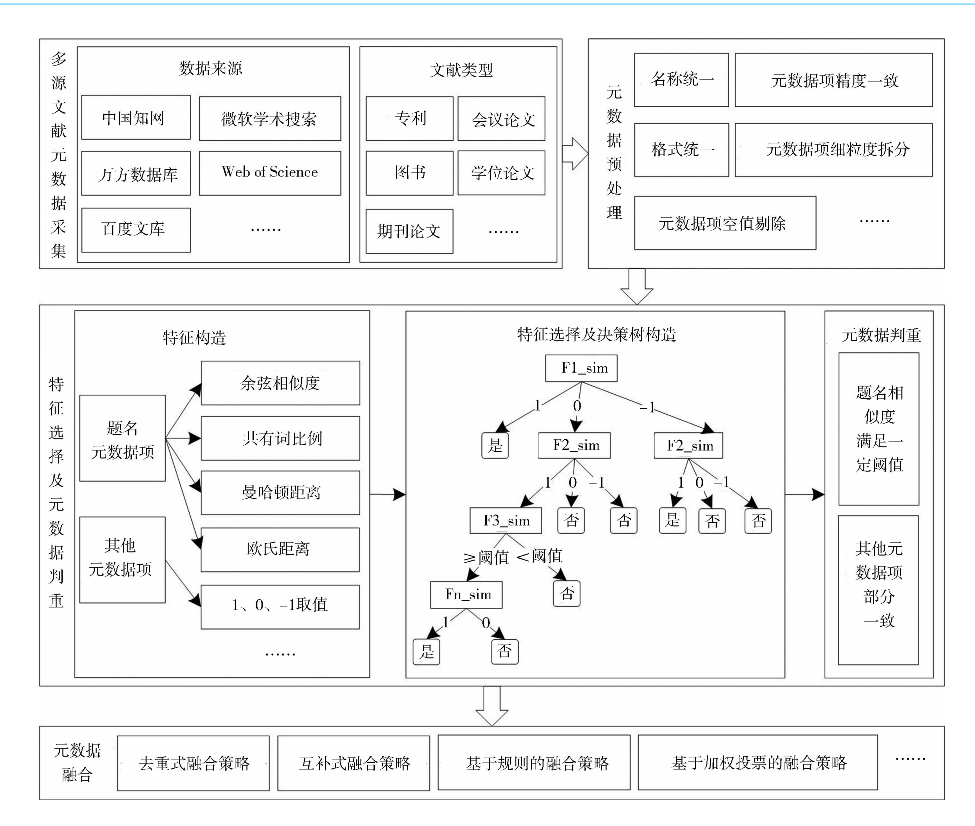


图 1　基于决策树的多源文献元数据融合模型

数据，如专利、学位论文、期刊论文、会议论文、图书等。此外，对于不同类型的文献，由于其包含的元数据项有差异，因此采集的元数据项也会有所不同。

3.2　元数据预处理

在元数据采集的基础上，需要对元数据进行预处理，即对各来源的元数据按照规定格式进行规范化处理，以便于后续元数据的判重与融合。在预处理环节需要遵循以下几点：①在规范设置时尽量选择各来源著录内容和格式的共同之处进行设计，需要注意的是，当两个元数据项名称不同但内容实为相同时，需要将该名称进行统一，如将“页码数”统一为“页数”。②尽量按照资源发现系统的著录标准保持精度一致（如时间、地点等），例如，在会议论文中，部分来源元数据著录的是会议地点的全称“中国福建厦门”，部分来源元数据则著录的是“厦门”，而资源发现系统的会议地点元数据著录标准要求保留局部即可，即在规范化处理时保留“厦门”。③元数据项细粒度拆分，将组合型的元数据项拆解到最细粒度，以保持统一，例如在湖北工业大学的图书元数据中，“出版地”“出版社”“出版时间”3 个元数据项被合并成一个元数据项“出版发行项”，以组合出现的元数据项会对元数据的判重与融合造成干扰，而具体的某一项或几项（如出版社）在判重与融合中具有一定的作用。因此，需要对该类元数据项按照内容进行拆解，才能与其他来源的相同元数据项进行判重及融合。④空值剔除，对于只有元数据项名称而无内容的空值项需要进行剔除，如会议论文中的“DOI”值全为空值，则将该项进行剔除。

3.3　特征选择及元数据判重

服务商需要确定哪些元数据纳入待判重元数据体系或待融合元数据体系，因此需对预处理后得到的规范元数据进行判重。由于原有方法主要是凭借经验制定相应的规则对元数据实施判重，自动化水平不高，效率较低；而且由于不同类型文献的元数据项不一样，其所用到的方法可能不同；此外对于同一种文献类型而言，不同来源的元数据项不一样，实际采用的方法可能也会有所区别，因此原方法的适应性较差。而对两条元数据判重的本质是通过对任意元数据项进行组合来判断是否为同一条元数据，答案只有是与否，因此本文将判重问题转化为分类问题，采用机器学习中的决策树算法，实现对元数据的自动化判重。依据元数据项构造出一系列特征，计算各个特征之间的相似度，基于相似度选择特征，在此基础上生成决策树，具体可分为以下几个步骤。

3.3.1　特征构造

依据元数据项构造特征，计算各个特征之间的相似度，特征类型不同，相似度计算方法有所差异，如余

李静，胡潜，李想，等. 基于决策树的多源文献元数据融合研究[J]. 图书情报工作,2022,66(6):118－125.

弦相似度、共有词比例、曼哈顿距离及 1、0、－1 离散取值等。由于元数据项存在不完整的情况，因此需要对数据进行分组训练，分组依据是两组数据至少有一项元数据项不为空，即两项至少有一项不为空。

(1)对于题名元数据项，由于各元数据提供方在著录时可能遵循的是不同的著录规则(如不著录题名副标题或不对题名上下标进行处理等)抑或是在著录时出错，均可能导致题名信息和长度存在不一致，因此，借鉴最短匹配的思想，采用共有词比例方法来计算题名相似度(见公式 1)[20－21]。

$$\mathrm{Sim}(M_a, N_b) = \frac{\mathrm{num}_{\mathrm{mutual}}(M_a, N_b)}{\mathrm{num}_{\mathrm{shorter}}(M_a, N_b)} \quad 公式(1)$$

式中，M_a 为元数据信息库 M 中的题名特征 a，N_b 为元数据信息库 N 中的题名特征 b；$\mathrm{Sim}(M_a, N_b)$ 为 M_a 和 N_b 的题名相似度；$\mathrm{num}_{\mathrm{mutual}}(M_a, N_b)$ 为 M_a 和 N_b 共有汉字或单词的总数，其中重复汉字或单词不做去重处理；$\mathrm{num}_{\mathrm{shorter}}(M_a, N_b)$ 为 M_a 和 N_b 中的最短长度特征的汉字或单词数。

需要关注的是，在计算题名相似度前应对题名序号进行一致性判断，原因在于部分文献资源(如图书)可能会存在"上册、中册、下册""(1)、(2)、(3)"及"续"等带有序号的信息，这类型文献资源相互之间具有较强的关联性，各元数据项的值均高度一致，易造成误判，因此当序号不一致时，判定题名特征相似度为 0，以此提高判重策略的准确性。

(2)对于其他类型的元数据项，由于只需要判断两项是否一致，并不需要判断一致的程度，因此，在构造特征时相似度取离散值，利用 1、0、－1 对特征值的相似度进行表示，1 表示两个特征一致，0 表示两个特征不一致，－1 表示两个特征其中一项为空，无法判断是否一致。

3.3.2 特征选择及决策树构造

决策树通常采用 ID3、C4.5、C5.0 和 CART 算法来构造[22－23]，由于 ID3 算法计算复杂度不高，输出结果易于理解，且针对离散型数据适应性较强，因此本文根据采集到的文献数据特点选择该算法构造决策树。ID3 算法的核心在于信息增益，因此将基于特征相似度计算出的信息增益定义为一个特征能够为分类系统带来多少信息量，即在一个条件下，信息复杂度减少的程度，信息量越大，说明该特征越重要，相应的信息增益越大，对整个系统复杂度减少的贡献越大[24]。信息增益能够表征数据集划分前后信息发生的变化，即引入特征 A 后原数据集 D 的不确定性减少的程度，见公式(2)、(3)、(4)。

$$infoGain(D \mid A) = H(D) - H(D|A) \quad 公式(2)$$

式中，infoGain(D|A)为信息增益，H(D)为 D 的信息熵，H(D|A)为给定 A 的条件下 D 的条件熵。

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|} \quad 公式(3)$$

式中，H(D)表示信息熵，|D|为数据样本总数，k 为特征个数，$|C_k|$为第 k 个特征的样本个数，$\frac{|C_k|}{|D|}$为随机变量的概率值。

$$H(D|A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} log_2 \frac{|D_{ik}|}{|D_i|}$$

公式(4)

式中，H(D|A)表示条件熵，D_i 为随机变量，$\frac{|D_i|}{|D|}$为选定特征的某个类别的概率，$|D_{ik}|$为在 Di 条件下某个类别的样本个数。

计算每个特征引入后的信息增益，获得信息增益最大的特征即为最优选择，依次选择特征，直至类别完全相同或是没有特征可进行选择，进而得到最终的决策树。

3.3.3 通过决策树对元数据判重

不同来源的文献在同一元数据项上可能产生不同值，仅凭单个元数据项作为判重条件可能会导致误判，因此需要识别出关于同一篇文献的多个元数据项并对其进行组合，才能进行元数据融合，否则可能引发新的数据错误。对不同来源但属于同一内容的元数据项实施判重，借鉴林鑫等[1]研究中对元数据的透视分析，本文采用的判重策略是题名相似度满足一定阈值，其他元数据项要求部分取值相等(如专利中的公开号、公开日、申请人等)，该策略具有一定的容错性，能够在保证准确率的同时兼顾召回率，避免对判重效果产生大的负面影响。

3.4 元数据融合

在对元数据实施判重后，即确定了各来源描述同一文献的元数据，接着对各元数据进行内容融合，提升元数据的整体质量，进而获取准确、一致、完整的文献资源。依据林鑫等[1]研究中对元数据的透视分析，本文针对不同元数据采用不同的元数据融合策略，具体为去重式融合策略、互补式融合策略、基于规则的融合策略和基于加权投票的融合策略。其中前两种融合策略只针对元数据本身，最为简单有效；后两种涉及到各来源元数据的质量问题和异构问题。在融合时，需要

针对不同情形，采取相应的融合策略。

3.4.1 去重式融合策略

多数情况下，各来源的元数据信息均较完整，关于同一文献资源的元数据非空且一致，此时仅需采取最简单的去重式融合策略，保留任意一个来源的元数据即可。如专利中的“题名”元数据项，各来源的元数据项均为“人工智能灭火机器人”，则直接将其作为融合后的“题名”元数据项。

3.4.2 互补式融合策略

多数元数据信息较为完整，但也存在部分元数据缺失的情况，因此需要对空值进行填补，此时采取互补性融合策略。通过判重策略确定为同一文献资源的各来源元数据，当只有某一来源的值非空时，则将其保留。如某一篇期刊论文中的“作者”元数据项，各来源中只有一个来源的“作者”项非空，其他来源的“作者”项均为空值，则将非空项信息进行保留，作为该文献的“作者”信息。互补式融合策略能够有效弥补空值对完整性带来的影响，虽实施简单但效果极其显著。

3.4.3 基于规则的融合策略

基于规则的融合策略主要是针对某一文献资源的各来源元数据均不为空，但只有一个来源的元数据著录标准符合规范，则对其进行保留。如各来源“题名”元数据项，当只有一个来源中的“题名”项包含副标题时，则将其作为该文献资源的“题名”信息。又如“题名”元数据项包含特殊字符，则将其切分为文字局部和特殊字符局部，保留符合规范的文字和特殊字符，将其作为“题名”信息。

3.4.4 基于加权投票的融合策略

基于加权投票的融合策略主要针对的是某一文献资源的各来源元数据均不为空，且符合著录规范的元数据有多个，此时则需要通过加权投票策略进行融合。针对每一条元数据信息，根据公式(5)对其权值进行计算，并将权值最大的作为最终结果。

$$W_j = \sum_{i=1}^{k} S_{i-j} \qquad \text{公式(5)}$$

式中，W_j 是指 j 元数据项的加权投票权重；S_{i-j} 是指 i 来源中 j 元数据项的质量得分，其分数来源于元数据透视环节的定量评价结果，视具体情况而定。

4 实验设计

为了对模型效果进行全面验证，选择样本数据时需要涵盖各类型文献资源，中国知网和万方数据库作为国内科研领域较常使用的文献资源数据库，资源覆盖范围广泛、内容包含全面，为科研学者提供了便捷的知识获取服务。因此，本文选择上述两个平台为文献资源数据库，选取专利、学位论文、期刊论文和会议论文为元数据对象来源；高校图书馆馆藏图书由于其学术价值较高，且考虑到数据的可获得性和获取便捷性，故图书元数据选自华中师范大学图书馆和湖北工业大学图书馆。

4.1 数据采集

笔者从中国知网和万方数据库中导出的专利元数据分别为 6 000 和 6 050 条，学位论文元数据分别为 5 358和 4 742 条，期刊论文元数据分别为 21 798 和 28 453条，会议论文元数据分别为 5 104 和 1 125 条；从华中师范大学图书馆和湖北工业大学图书馆中导出的图书元数据分别为 2 488 和 2 484 条。具体的元数据如表 1 所示：

表 1 多文献类型各来源元数据项

文献类型	数据来源	元数据项
专利	中国知网	题名/专利名称、公开号、申请人、申请机构、作者/发明人、申请日、公开日、国省名称、专利类别名称、摘要、主权项、数据库、中图分类号、ISSN
	万方数据库	题名、申请/专利号、公开/公告号、申请人、发明/设计人、申请日期、公开/公告日、主权项、摘要
学位论文	中国知网	标题、作者、关键词、机构、摘要、专辑、专题、分类号、指导老师
	万方数据库	标题、摘要、DOI、关键词、作者、学位授予单位、授予学位、学科专业、导师姓名、学位年度、语种、分类号、出版时间
期刊论文	中国知网	刊名、标题、作者、关键词、机构、摘要、卷、期、年、页数、页码
	万方数据库	标题、摘要、DOI、关键词、作者、作者单位、刊名、年、卷、期、所属期刊栏目、分类号、出版时间、页数、页码
会议论文	中国知网	标题、作者、关键词、作者单位、摘要、基金、会议名称、会议时间、会议地点、专题、专辑、分类号、页码、页数
	万方数据库	标题、摘要、DOI、关键词、作者、会议名称、作者单位、母体文献、会议时间、会议地点、语种、分类码、页码
图书	华中师范大学	MARC 号、索书号、题名、责任者、出版社、出版年、标准号、文献类型
	湖北工业大学	题名、责任者、出版发行地、出版社、出版时间、定价、载体形态项、学科主题、分类号、提要文摘附注

利用实验数据对模型进行验证，对于判重策略的效果验证采用的是准确率和召回率；对于融合效果的验证采用的是元数据项质量提升比例。判重策略的准确率和召回率的计算如公式(6)和公式(7)所示：

$$\text{Precision} = \frac{T}{m} * 100\% \qquad 公式(6)$$

式中,T 为判重策略识别出的篇数且实际重复的篇数;m 为判重策略识别出的篇数。

$$\text{Recall} = \frac{T}{T+n} * 100\% \qquad 公式(7)$$

式中,n 为判重策略未识别出的篇数。

4.2　元数据判重与融合

首先是对采集到的 5 种类型元数据进行预处理,在实施过程中主要包含几个部分:对元数据名称和格式进行统一化、保持元数据项精度一致、将元数据项进行细粒度拆分及对元数据项空值进行剔除。

其次是特征构造,利用元数据项构造特征并计算特征相似度,所使用的元数据项要求至少有一边非空,即至少有一项有取值,以计算相似度。对于题名特征,采用共有词比例方法得到题名相似度,其他特征采用 1、0、-1 离散值对相似度进行表示,两者一致为 1,两者不一致为 0,两者有一边为空用 -1 表示。各来源文献元数据判重及融合实际所选用的元数据项如表 2 所示:

表 2　各来源文献元数据判重及融合实际选用元数据项

文献类型	判重使用的元数据项	融合使用的元数据项
专利	题名/专利名称、公开号、申请人、作者/发明人、申请日、公开日	题名/专利名称、公开号、申请人、作者/发明人、申请日、公开日、申请机构、国省名称、专利类别名称、申请/专利号、摘要、主权项
学位论文	标题、作者、关键词、机构、分类号、指导老师	标题、作者、关键词、机构、分类号、指导老师、摘要、专辑、专题、DOI、授予学位、学科专业
期刊论文	标题、作者、关键词、机构、期刊、卷、期、年、页数、页码	标题、作者、关键词、机构、期刊、卷、期、年、页数、页码、摘要、DOI、所属期刊栏目、分类号
会议论文	标题、作者、关键词、作者单位、会议名称、会议时间、会议地点、分类号、页码	标题、作者、关键词、作者单位、会议名称、会议时间、会议地点、分类号、页码、摘要、基金、专题、专辑、页数、母体文献
图书	题名、责任者、出版社、出版年、ISBN	题名、责任者、出版社、出版年、ISBN、出版地、定价、页数、载体形态项

首先,为确定题名相似度阈值,从中国知网和万方数据库中各抽取专利、期刊论文、学术论文和会议论文的元数据 1 000 条,从两个图书馆数据库中抽取图书元数据 1 000 条,共计 5 000 条元数据作为训练集,分别对两个来源的"题名"元数据进行相似度计算,当相似度阈值为 0.85 时,5 种文献类型的元数据判重策略的准确率和召回率均达到了 99% 以上,判重效果显著。综合 5 种类型文献元数据,故将题名相似度阈值设为 0.85。

其次,根据特征相似度进行特征选择及决策树构造。基于特征相似度计算出的信息增益越大,对于各来源文献资源元数据判重的意义越大,优先选择信息增益最大的特征作为根节点,递归构造决策树。以图书为例构造决策树,详如图 2 所示:

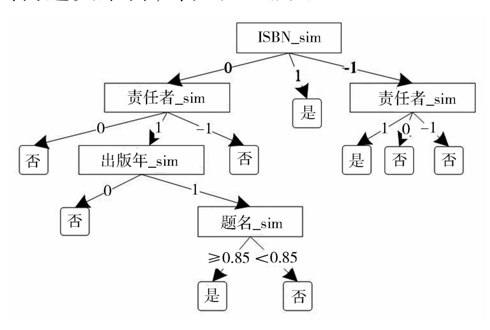


图 2　图书决策树

通过对元数据的分析,在满足 2/3 元数据项重复的条件下,将其判定为重复元数据,并生成重复元数据集合,在此基础上,进行元数据融合。具体包含以下情形:①当两个来源的元数据项均非空且取值一致时,采取去重式融合策略得到融合结果;②当两个来源的元数据项均不为空且取值不一致,但只有其中一项符合著录规范,则采用基于规则的融合策略,将规范性的元数据作为融合结果;③当两个来源的元数据项均不为空且取值不一致,但均符合著录规范时,则采取基于加权投票的策略,引入其他来源对其进行投票计算,选择权值较大的作为融合结果;④当两个来源的元数据项其中有一个为空时,采取互补式融合策略,将其作为融合结果。最后对各融合结果进行重组,生成对于文献资源准确、一致且完整的描述。

4.3　实验结果评价

本文从元数据判重的准确率、召回率和融合后元数据项质量提升比例两个方面对多源元数据融合模型进行效果检验。

首先是对元数据判重效果进行评价,从 5 种类型的元数据判重结果中各随机抽取 500 条,共计 2 500 条作为测试集计算准确率和召回率,结果如表 3 所示:

表 3　各类型文献元数据判重策略效果

	专利/%	学位论文/%	期刊论文/%	会议论文/%	图书/%
准确率	100	100	99.3	100	100
召回率	100	99.1	98.7	99.5	100

从表 3 看出,5 种文献类型元数据判重策略的准确率均达到了 99% 以上,召回率达到了 98% 以上,总体效果较好,能够验证模型中判重策略的合理性。

其次通过元数据项质量提升比例对元数据融合效果进行评价。从通过判重策略并实施融合后的文献元数据中各抽取 500 条,对元数据项质量提升比例进行统计分析,具体如表 4 所示:

表 4　各类型文献元数据融合策略效果

	专利 /%	学位论文 /%	期刊论文 /%	会议论文 /%	图书 /%
元数据项质量提升比例	15.15%	36.80%	15.29%	52.63%	15.38%
元数据项质量不变比例	84.85%	63.20%	84.71%	47.37%	84.62%
元数据项质量降低比例	0	0	0	0	0

从表 4 可以看出,经过融合策略后,各类型文献的元数据项质量均得到了一定程度的提升,其中,会议论文的提升幅度最大,达到 52.63%,原因在于万方数据库中部分元数据项的空值情况较多(如页码、作者单位等),通过互补式融合策略能够对其进行有效补充;其次是学位论文,提升比例为 36.80%,主要在于万方数据库未按照学位论文的原文关键词对其进行元数据项标注,通过基于加权投票的融合策略能够提升其准确性;其他 3 种类型文献的元数据项质量提升比例均在 15% 左右,原因在于各元数据项的准确性和完整性较高,因此元数据项质量提升比例不大。

5　结语

为了更好地利用元数据资源,促进各类型文献资源的开放融合与协作共享,优化用户的资源发现服务体验,本文针对元数据存在的质量问题设计了基于决策树的多源文献元数据融合模型,对元数据实施判重与融合,从经验为主向自动化转变,扩大了模型适用范围,并以中国知网、万方数据库、华中师范大学图书馆及湖北工业大学图书馆的各类型元数据为例进行效果验证。实验结果表明,该策略对各类型文献元数据的融合实现具有良好效果,且在保障效果的前提之下,提升了整个过程的自动化水平,效率更高。但本研究还存在一些问题有待后续改进,主要包括:①针对各类型元数据,均只选取了两个来源的中文文献对模型进行验证,在元数据融合时,无法对基于加权的内容融合策略进行效果验证;②仅针对中文文献资源元数据进行了处理,对于多语言文献资源元数据的融合并未进行验证,后续应针对该类情况优化融合模型,以增强模型的普适性。

参考文献:

[1] 林鑫,李想,李静. 资源发现系统中基于多源数据融合的文献元数据质量提升[J]. 情报理论与实践,2021,44(5):122-126,186.
[2] PARK J R, TOSAKA Y. Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms[J]. Cataloging & classification quarterly,2010,48(8):696-715.
[3] STVILIA B, TWIDALE M B, SMITH L C, et al. Information quality work organization in Wikipedia[J]. Journal of the American Society for Information Science and Technology,2008,59(6):983-1001.
[4] BRUCE T R, HILLMANN D I. The continuum of metadata quality: defining, expressing, exploiting[C]//HILLMANN D I, WEATBROOKS E L. Metadata in practice. Chicago: American Library Association,2004:238-256.
[5] 黄莺,李建阳. 元数据质量评估方法及模型研究[J]. 图书馆学研究,2013,(12):52-56.
[6] 翟军,陶晨阳,李晓彤. 开放政府数据质量评估研究进展及启示[J]. 图书馆,2018(12):74-79.
[7] 黄刚,袁满,吴秀英. 元数据驱动的数据质量评估体系架构研究[J]. 计算机工程与应用,2013,49(8):114-119,181.
[8] 张晓娟,谭婧. 我国省级政府数据开放平台元数据质量评估研究[J]. 电子政务,2019(3):58-71.
[9] 董微,赵捷. 开放期刊资源元数据质量管理研究[J]. 中国科技资源导刊,2018,50(3):82-86.
[10] 刘家真,廖茹. 电子文件管理元数据的质量控制与管理[J]. 图书情报知识,2009(6):91-96,102.
[11] 寇晶晶,贾君枝. 高校图书馆资源发现系统中文检索性能比较分析[J]. 国家图书馆学刊,2016,25(6):71-79.
[12] LI G L, WANG J N, ZHENG Y D, et al. Crowdsourced data management: a survey[J]. IEEE transactions on knowledge and data engineering, 2016,28(9):2296-2319.
[13] 李慧佳,马建玲,张秀秀,等. 元数据语义化映射过程研究——以中科院机构名称规范控制库为例[J]. 图书馆论坛,2017,37(12):72-79.
[14] 孙锐,杨新涯,魏群义,等. 文献资产元数据仓储建设关键问题研究——以重庆大学图书馆为例[J]. 大学图书馆学报,2018,36(2):18-24.
[15] 鲁丹,李欣. 数字人文环境下异构方志元数据整合策略[J]. 图书馆论坛,2019,39(4):158-165.
[16] MANGUINHAS H, JOSE B. Quality control of metadata: a case with UNIMARC[J]. ECDL,2006,4172(3):244-255.
[17] 曹月珍,马建玲. 国内外元数据质量控制的研究进展与发展趋势[J]. 图书与情报,2013(6):101-104.
[18] 王利亚,邱航,陈若雅. 基于元数据可追溯性的健康医疗大数据治理方法及可视化呈现[J]. 中国卫生信息管理杂志,2019,16(6):661-666.
[19] 严承希,房小可. 开放世界视角:面向多源词表的知识融合框架

MtFFO 研究[J]. 中国图书馆学报,2017,43(4):114－129.
[20] 储光,胡学钢,张玉红. 基于语义的文本数据流概念漂移检测算法[J]. 计算机工程,2018,44(2):24－30.
[21] 李静,胡潜. 多语言 UGC 环境下 MOOC 课程笔记自动生成[J]. 情报理论与实践,2021,44(11):173－179.
[22] 唐亮,李飞. 基于决策树的车联网安全态势预测模型研究[J]. 计算机科学,2021,48(S1):514－517.
[23] 李勇男. 信息增益决策树在反恐情报分析中的应用研究[J]. 情报科学,2018,36(4):80－84,149.
[24] 吴鹏,肖维聪,楚榕珍. 基于模型检测的财经舆情可信度研究[J]. 情报学报,2020,39(6):619－629.

作者贡献说明:
李静:负责论文撰写与实验实施;
胡潜:负责论文选题与论文修订;
李想:参与实验实施与文献资料搜集;
肖兵:参与实验实施与数据分析。

Research on Metadata Fusion of Multi-Source Documents Based on the Decision Tree

Li Jing Hu Qian Li Xiang Xiao Bing
School of Information Management of Central Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] Constructing a multi-source document metadata fusion model will help improve the overall quality of document metadata, promote metadata management and utilization in the resource discovery system, and optimize user resource discovery service experience. In view of the document metadata duplication judgment strategy proposed by the writers before, this paper optimizes the strategy from experience-oriented to automated, and improves the automation level in the whole process on the premise of guaranteeing the duplication judgment and fusion effect. [**Method/process**] The metadata items of different types of documents were different, and the metadata items of the same document from different sources were different, which will make the method of judging duplication different. An automatic multi-source document metadata fusion model based on the decision tree was proposed, which transformed a duplication judgment problem into a classification problem. This paper selected features according to feature similarity and constructed the decision tree, on this basis, it implemented metadata duplication judgment and fusion, and took different types of document resource metadata as examples to conduct experiments to verify the effectiveness of the strategy. [**Result/conclusion**] The results show that for the five document types of metadata, the accuracy of the duplication judgment strategy is more than 99%, and the recall rate is more than 98%. The overall effect is good. Judgment on the effect of the fusion strategy, the quality improvement ratios of the metadata items of patents, dissertations, journal papers, conference papers and books are 15.15%, 36.80%, 15.29%, 52.63% and 15.38% respectively, all of which have significant improvement.

Keywords: multi-source metadata the decision tree metadata duplication judgment metadata fusion